Earth System Grid Federation (ESGF):
(ESGF: Peer-to-Peer)

How To Build
An Elastic Distributed System
Over "Big Data"

ESGF P2P:
Gavin M. Bell
May 10, 2011

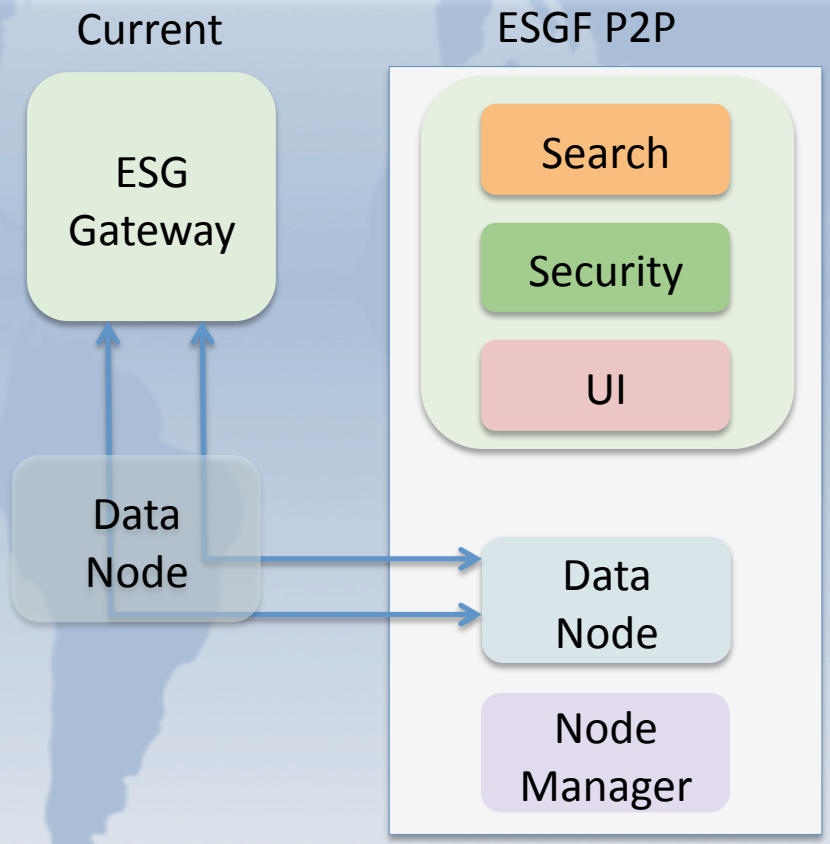
Overview

- The **Earth System Grid *Federation*** (**ESG*F*.org**) is indeed a spontaneous, unfunded, community-driven, open-source, collaborative effort to address the task of managing large amounts of distributed scientific data. Doing so requires dealing with the challenges of data distribution, data management and software design.

- **Problem**: We have lots and lots of data… "Big Data" (~1PB++)
  - How do we manage it all?
  - How do we store it, find it, access it, manipulate it?
- **Solution**:
  - Build an *elastic* distributed system over this "Big Data"
  - Create distributed architecture that scales horizontally
  - Build a system that is flexible and resilient to perturbations.
- **Reason**:
  - Give scientists the tools to enable even Bigger Science

- **The ESGF P2P Architecture:** "From Many One Data*space*"

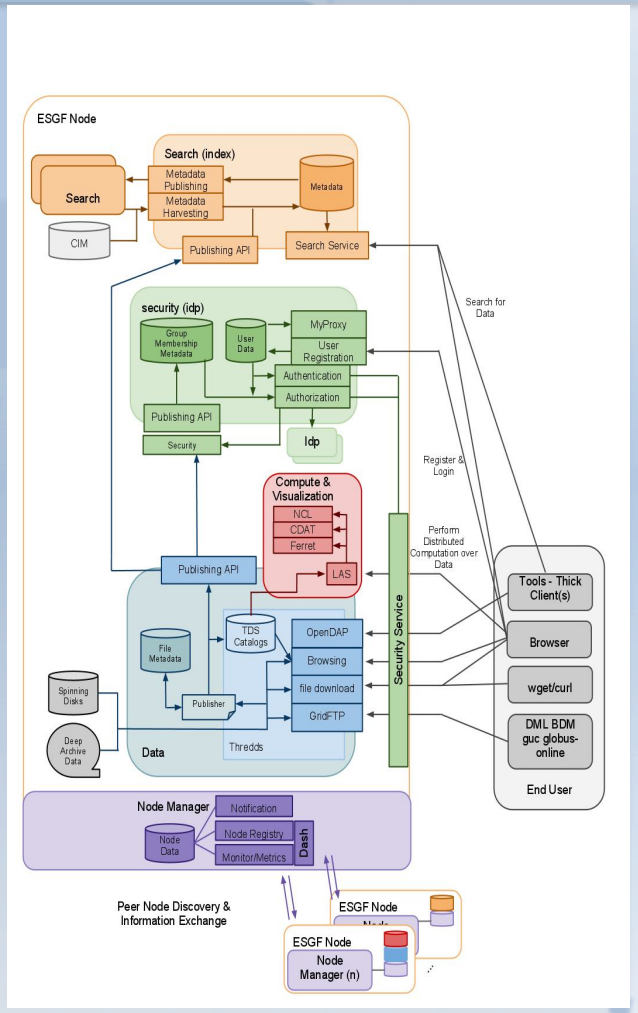
ESGF Development



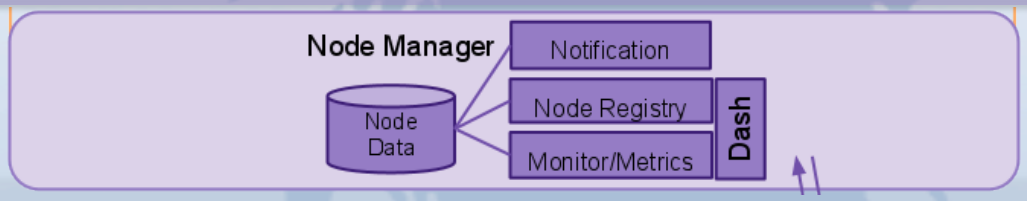
- Component Architecture
- Extensibility
- Enhanced Metadata Display
- Full support for Observations
- UI Options for Notification
- Integrated Security
- Additional Products
- Geospatial & Temporal Search
- Automatic Metadata Replication
- Enhanced scalability
- Dashboard showing Metrics
- Secure OPeNDAP access
- Rich analysis climate support

ESGF Node Architecture

- Nodes currently come in 4 configurations (any or all may co-exist simultaneously)
  - "**Index**" Provides searching and indexing across the space. It allows for robust and flexible distributed searching that is free text and facet driven.
  - "**IDP**" **(security)** Provides user account and credential information management. Along with the security framework we are able to facilitate seamless SSO access to dataspace holdings.
  - "**Compute**" Provides tools for doing distributed computation on data in the dataspace including visualizations,
  - "**Data**" Provides direct access to data, maintains the associated meta data and **publishing**. The publisher is the central component that scans data holdings, organizes the metadata and posts the metadata for indexing.
- There is an **installer** that will take a machine from tabula rasa to fully configured. (*nix)

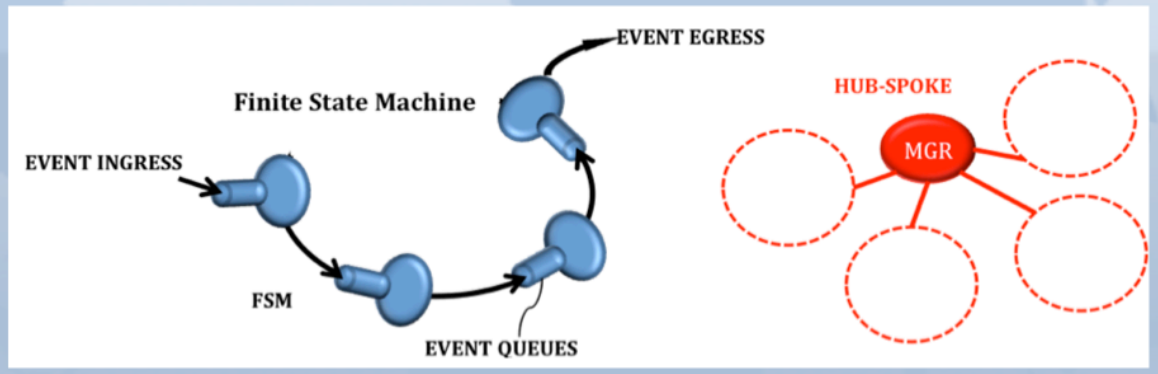


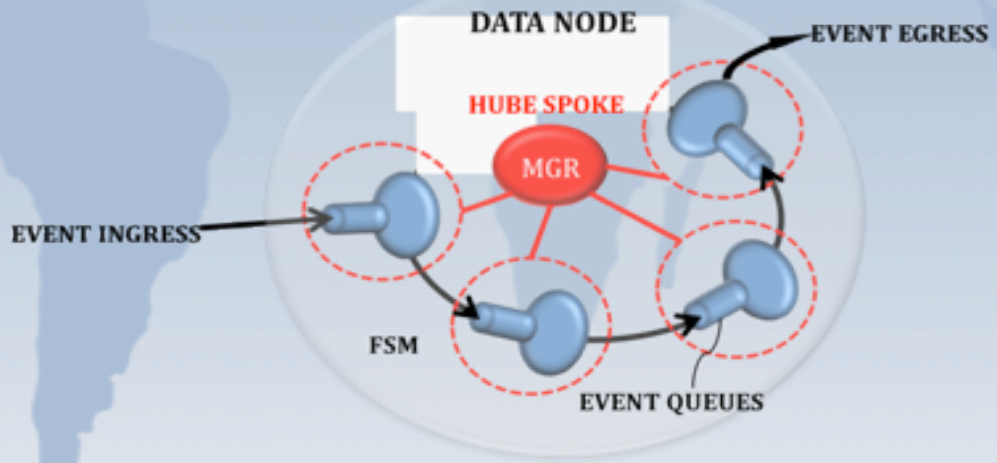
ESGF Node Manager



- **Node Manager**
  - Consistent component across all configurations.
  - The underlying coordinating entity for inter and intra node coordination.
  - Built on a lightweight event passing model
  - Currently supports **Metrics**, **Monitoring**, **Notification** and **Registry** as the core components.
  - Has its own front-end application coming online called the **Dashboard**.
- **Event Driven**
  - Intra-Node: Events are passed through to components connected in a FSM-like configuration.
  - Inter-Node: Events are also passed among nodes to facilitate information exchange etc.
  - Under this event driven model, tasks are processed as a FSM where “state transitions” are triggered by events.
  - Limits the complexity of building large systems and addresses the inherent difficulties of adapting to load in a purely thread-based model, while still supporting massive concurrency. The node manager manages message queues and can adaptively modify the FSM to address performance degradation.
  - To compliment the FSM event passing model - a “hub-spoke” management model.

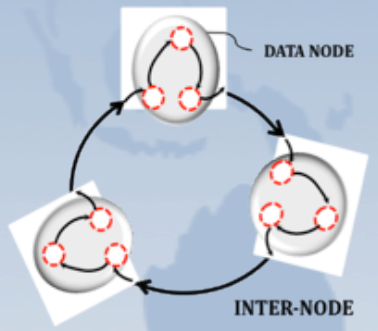
ESGF Node Manager





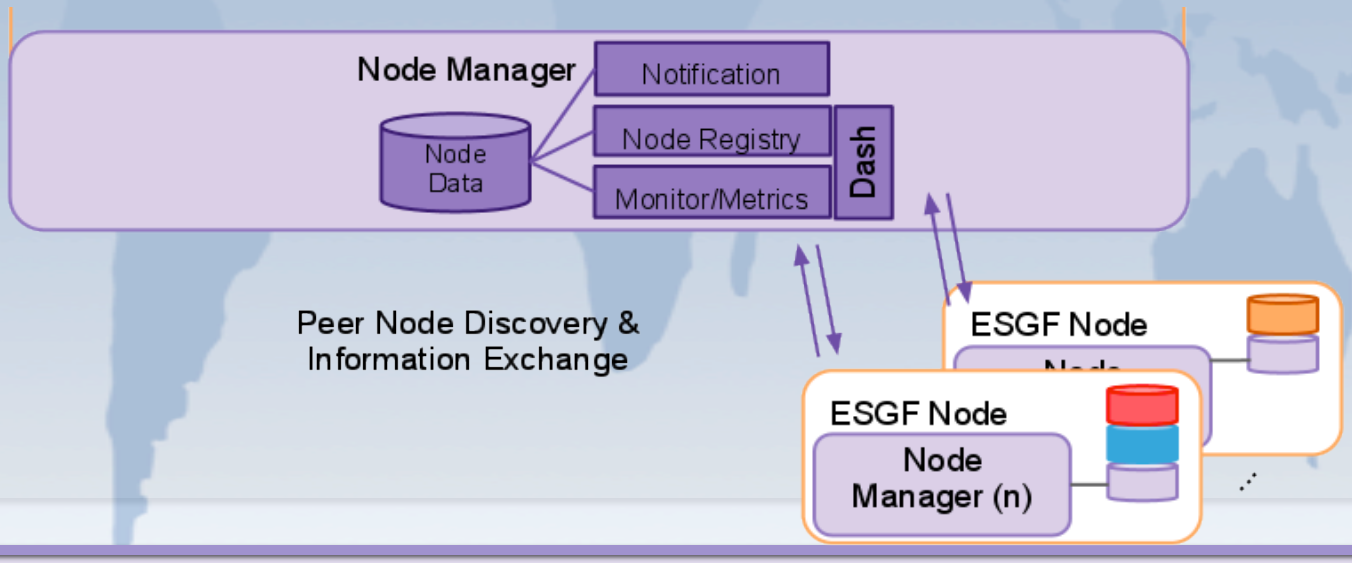
ESGF Node Manager

- **Gossip Protocol**
  - The Node Manager implements a gossip protocol to maintain a peer to peer global state.
  - This provides an eventually consistent model of participating peers.
  - Efficient protocol providing exponential information distribution in O (log n) rounds.



ESGF Node Manager

- **Elasticity**
  - The afore mentioned architecture, components and protocol, *together* create an *elastic* network that is adaptive and flexible enough to meet our current data distribution needs and we think future needs as well.
  - Nodes join the ESGF automatically and then quickly learn the presence of other nodes in the space and form an overly mesh network. The Node Manager's *Registry* component maintains the node's local state of the network by gossiping registration state. The state information consists of the available services and functionality that the node is providing at any given time. (more on this in two slides)
- Peer-To-Peer coordination among ESGF Nodes creates a vast *dataspace* for us to do science over.



Core Components

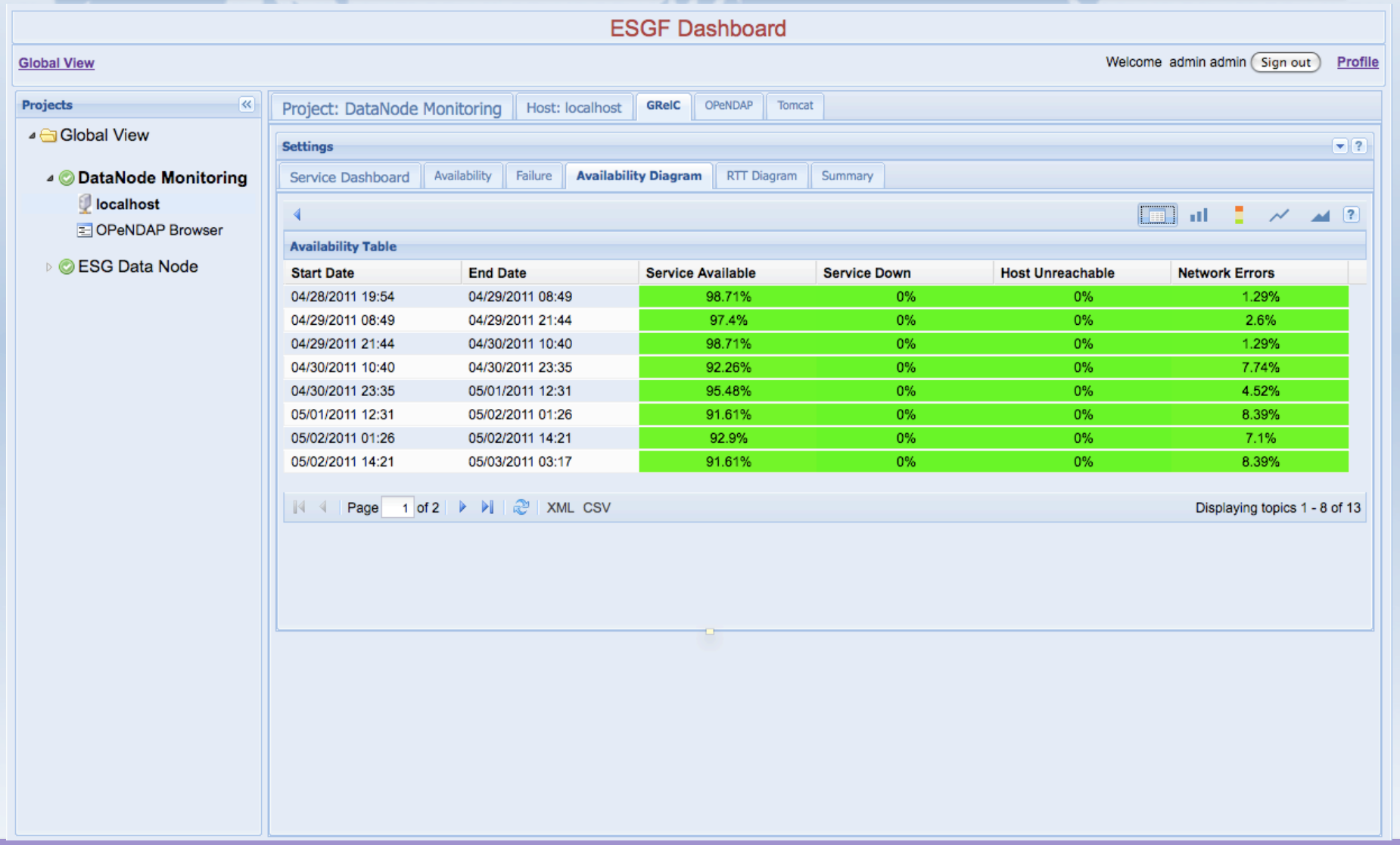
- **Monitoring**
  - Monitors system performance statistics
  - Disk space
  - CPU performance and uptime
  - Transfer rates
- **Metrics**
  - Provides information regarding user interaction with the node, specifically file download statistics. Who? What? Where? When?
- **Notification**
  - When published datasets are updated/changed notification emails are sent out to users for which the metrics data indicates have the old data and should be notified.
  - Email is the currently integrated mode but the architecture would support any number of possible notification modes.
- **Registry**
  - The central component for maintaining a node's world view of its peers. This information is gossiped and updated accordingly. Derivative representations such as application specific white lists are generated from this global view information. (Ex LAS sisters, IDP nodes)
- **Dashboard**
  - New web front-end to the Node Manager coming on line shortly that will provide a interface into the node's monitoring, metrics, notification and registry information.

The Registry

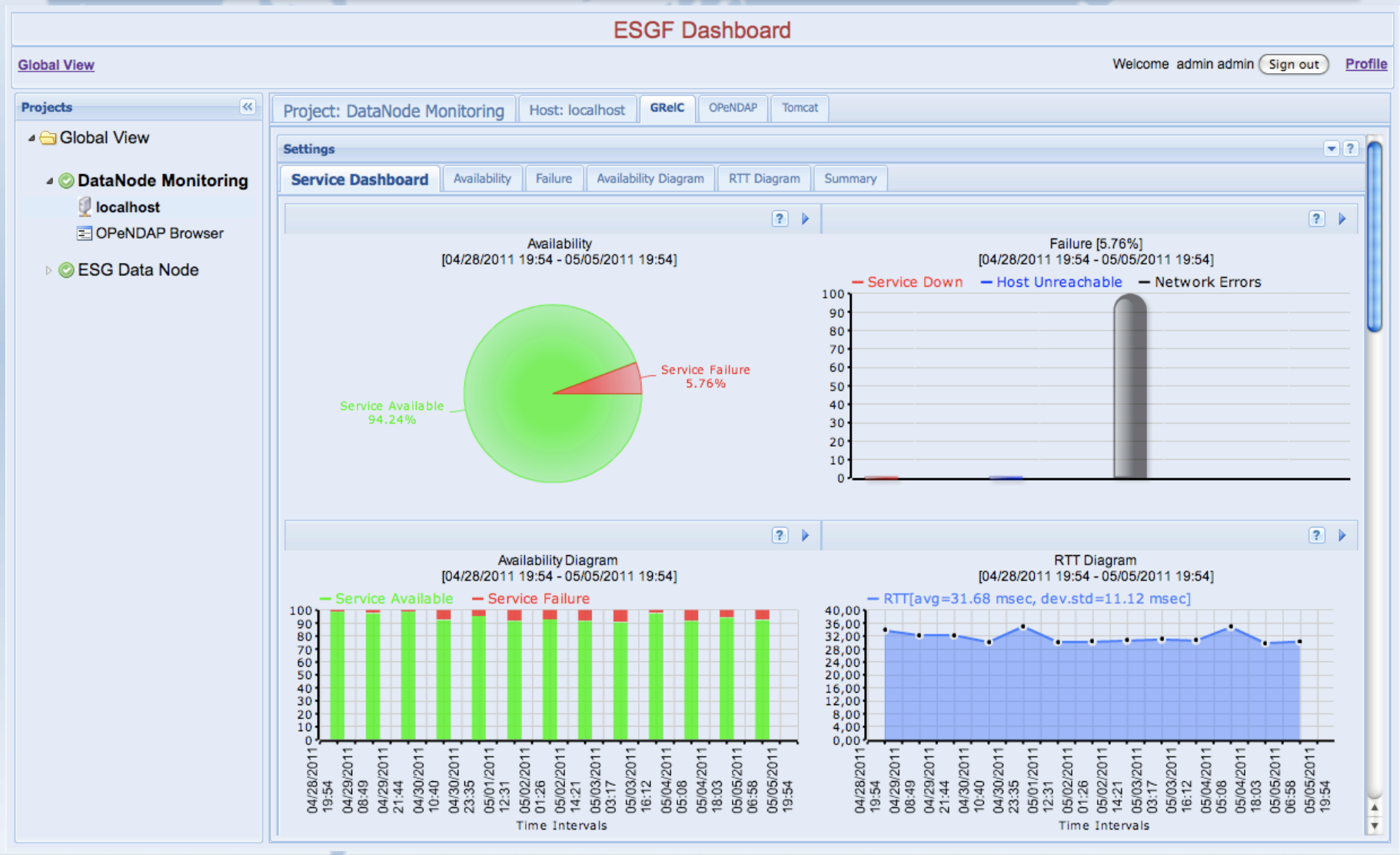
- Houses the local representation of the global “state” of the P2P mesh network.
- The state is described by an XML schema that generates a “registration” document (located in a web reachable location on a node)
- Anatomy of the Registration document.
  - Consists of “node” elements – one for each node in the network
  - Describes the services provided by a node along with relevant info.
    - What services are present
    - What is the service URL/URI endpoint
    - What port does it listen to (ex: GridFTP)
    - What is the version of the service
    - What is the certificate of that node
    - What groups are supported
    - What configuration is currently being advertized. (data, index, …)
- This information is gossiped and is eventually consistent (“BASE” strategy)
- Forms basis for derived information used by ORP, LAS, SEARCH
- ~~Directory Service~~ we call it…, The Registry

Dashboard Sneak Peek

Dashboard Sneak Peek

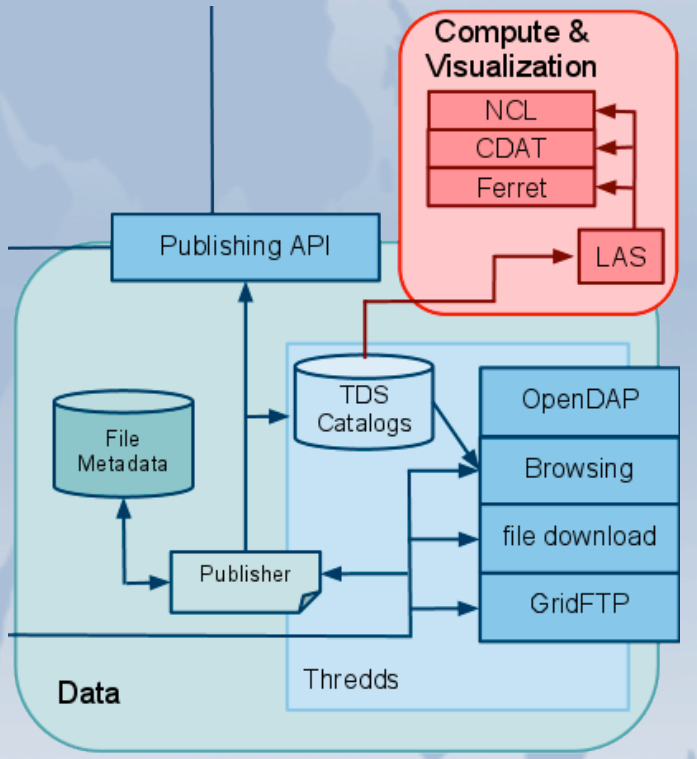


Dashboard Sneak Peek

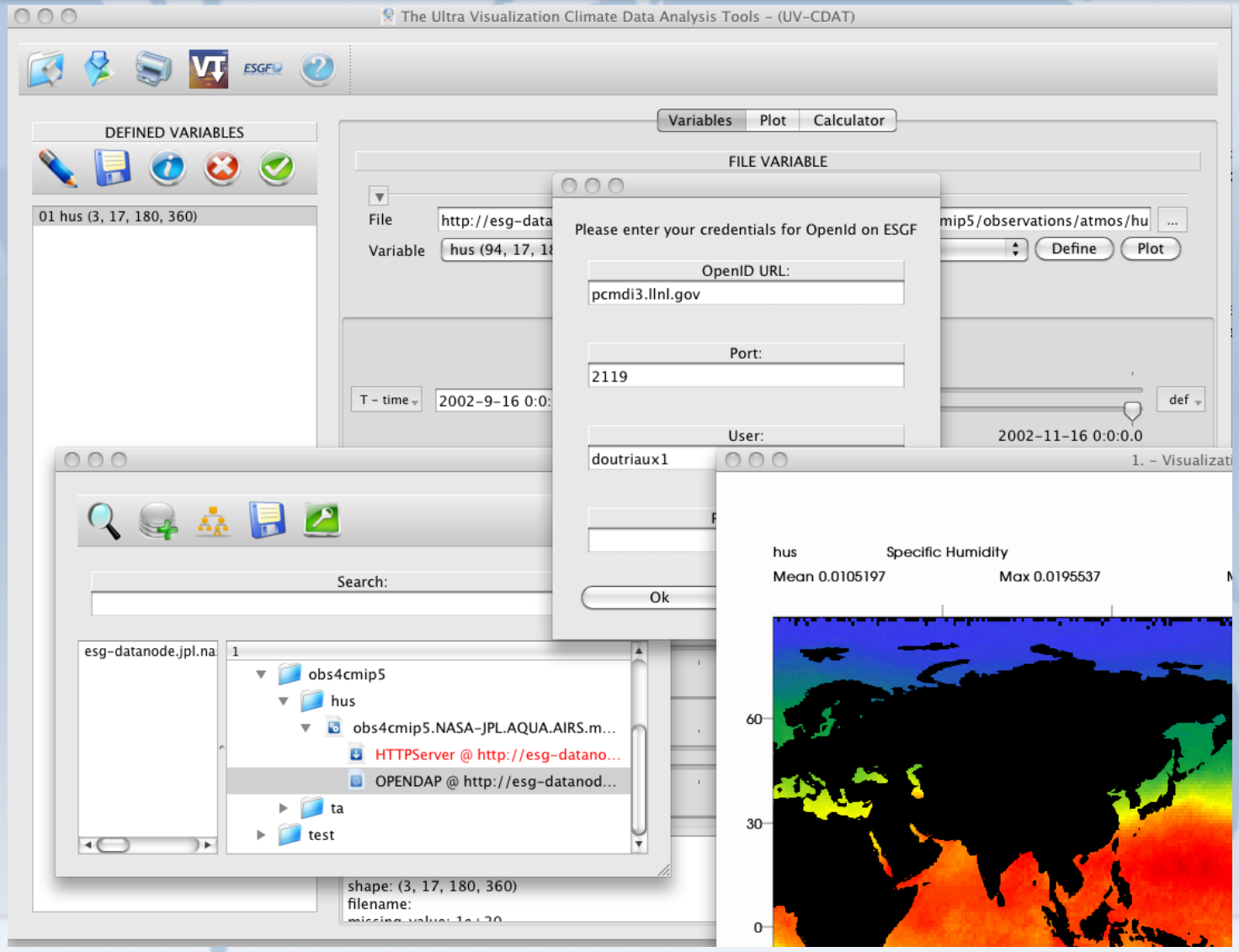


Data & Compute Node Types

- Under the "***Data***" type configuration of the ESGF Node we install the **Publisher**, the center piece of the "Data" configuration, with its *newly updated GUI* as well as **Thredds** and **Globus' GridFTP** for file transfer. We are currently integrating support for **Globus OnLine (GO)** to facilitate large-scale data downloads and replication.
- Under the "***Compute***" type configuration we also install the **Live Access Server (LAS) Confluence Server** for visualization and analysis.
- It is also important that we make it easy for additional clients to connect into the ESGF dataspace. Rich, non-web based clients such as **UV-CDAT** is able to connect directly to the ESGF P2P Node and search and browse for data and directly perform computation said data. (This will be demonstrated during the demonstration session).



UV-CDAT + ESGF P2P



http://esgf.org

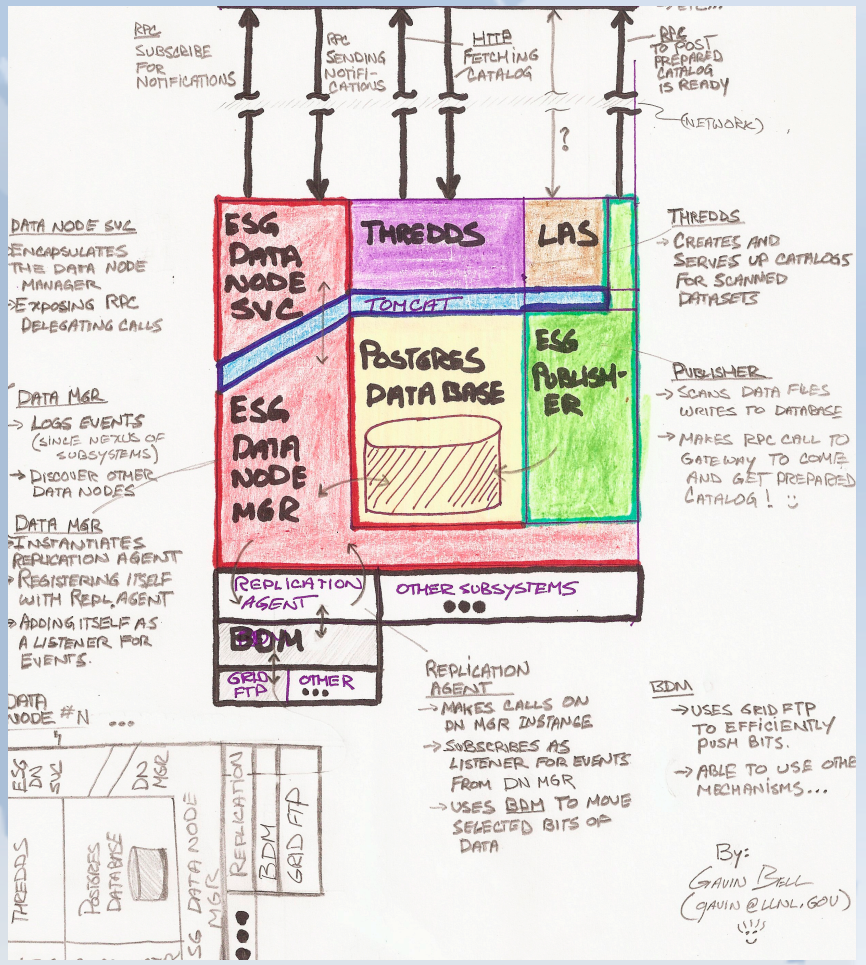
- **ESGF.org is the home of this open source effort. Here we maintain the infrastructure and tools that facilitates communication, collaboration and coding!**
- **Currently hosting 12 projects.**
- **The usual suspects of tools…**
  - Site & Project Sites
  - Bugzilla
  - Git
  - Wiki
  - Blog
  - Mailing Lists (esg-node-{dev,user}@lists.llnl.gov, etc…)
  - Artifactory

Questions?

- ESGF.org effort is just under 10 months old, started from humble beginnings. Our progress to date is a testament to the many talented people across institutions working together, collaborating, coordinating and coding!
- Join in, it's fun… **promise**!

"Never mistake a clear view for a short distance"

paul saffo.



ESGF.org P2P Team

- Gavin M. Bell, Bob Drach, Charles Doutriaux, Renata McCoy, Dean Williams [LLNL/PCMDI]
- Luca Cinquini, Dan Crichton, Chris Mattmann [NASA/JPL]
- John Harney, Galen Shipman, Feiyi Wang, [ORNL]
- Roland Schweitzer [NOAA/PMEL]
- Rachana Ananthakrishnan, Neill Miller [ANL]
- Estani Gonzales, Stephen Kindermann [DKRZ]
- Philip Kershaw, Stephen Pascoe [BADC]
- Luca Cinquini, Cecelia DeLuca, Sylvia Murphy, [NOAA/ESRL]
- Sandro Fiore, Giovanni Aloisio [Salento Univ.]

et. al